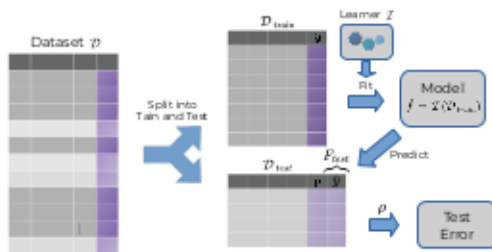


# TEST ERROR AND HOLD-OUT SPLITTING

- Simulate prediction on unseen data, to avoid optimistic bias:

$$\rho(\mathbf{y}_{\text{test}}, \mathbf{F}_{\text{test}}) \text{ where } \mathbf{F}_{\text{test}} = \begin{bmatrix} \hat{f}_{\mathcal{D}_{\text{train}}}(\mathbf{x}_{\text{test}}^{(1)}) \\ \vdots \\ \hat{f}_{\mathcal{D}_{\text{train}}}(\mathbf{x}_{\text{test}}^{(m)}) \end{bmatrix}$$

- Partition data, e.g., 2/3 for train and 1/3 for test.

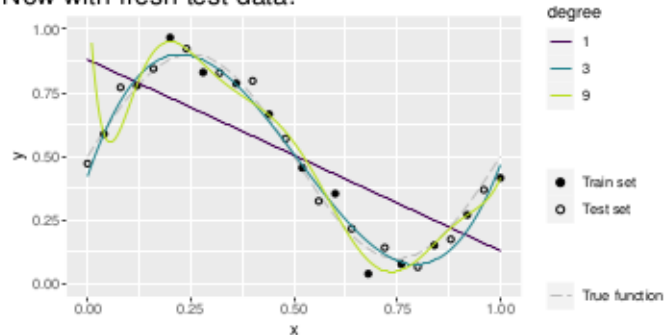


A.k.a. holdout splitting.



## EXAMPLE: POLYNOMIAL REGRESSION / 2

Now with fresh test data:



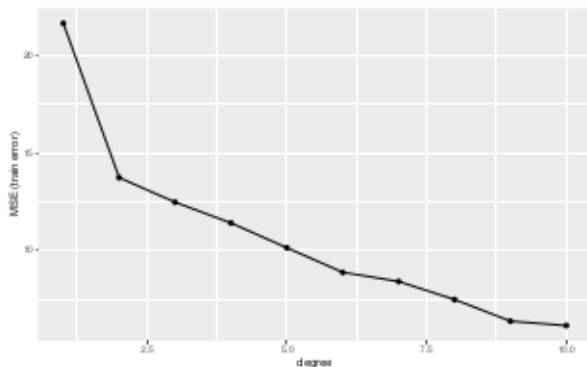
- $d = 1$ : MSE = 0.038: clearly underfitting
- $d = 3$ : MSE = 0.002: pretty OK
- $d = 9$ : MSE = 0.046: clearly overfitting

While train error monotonically decreases in  $d$ , test error shows that high- $d$  polynomials overfit.

## TRAINING VS. TEST ERROR / 2

### The training error...

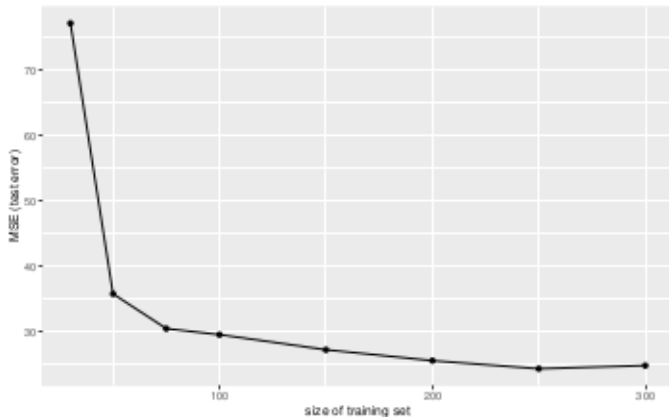
- decreases with increasing model complexity as the model gets better at learning more complex structures.



## TRAINING VS. TEST ERROR / 3

### The test error...

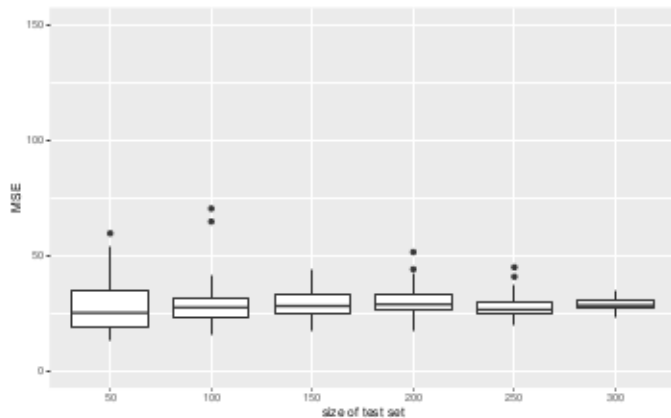
- will typically decrease with larger training set size as the model generalizes better with more data to learn from.



## TRAINING VS. TEST ERROR / 4

### The test error...

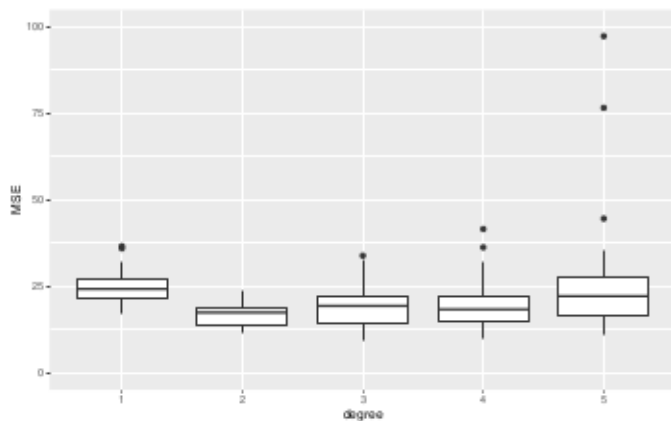
- will have higher variance with smaller test set size.



## TRAINING VS. TEST ERROR / 5

### The test error...

- will have higher variance with increasing model complexity.

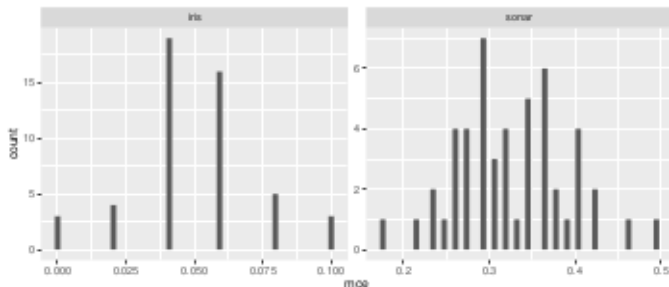


# BIAS AND VARIANCE

- Test error is a good estimator of GE, given a) we have enough data b) test data is representative i.i.d.
- Estimates for smaller test sets can fluctuate considerably – this is why we use resampling in such situations.

Repeated  $\frac{2}{3} / \frac{1}{3}$  holdout splits:

*iris* ( $n = 150$ ) and *sonar* ( $n = 208$ ).



## BIAS-VARIANCE OF HOLD-OUT – EXPERIMENT

Hold-out sampling produces a trade-off between **bias** and **variance** that is controlled by split ratio.

- Smaller training set  $\rightarrow$  poor fit, pessimistic bias in  $\widehat{GE}$ .
- Smaller test set  $\rightarrow$  high variance.

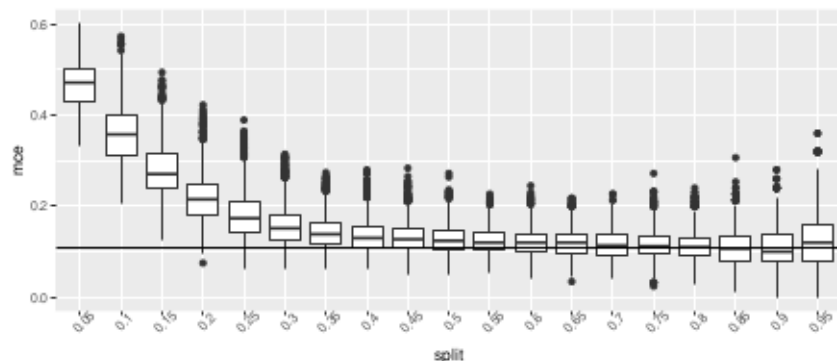


Experiment:

- `spirals` data ( $sd = 0.1$ ), with CART tree.
- Goal: estimate real performance of a model with  $|\mathcal{D}_{\text{train}}| = 500$ .
- Split rates  $s \in \{0.05, 0.10, \dots, 0.95\}$  with  $|\mathcal{D}_{\text{train}}| = s \cdot 500$ .
- Estimate error on  $\mathcal{D}_{\text{test}}$  with  $|\mathcal{D}_{\text{test}}| = (1 - s) \cdot 500$ .
- 50 repeats for each split rate.
- Get "true" performance by often sampling 500 points, fit learner, then eval on  $10^5$  fresh points.



## BIAS-VARIANCE OF HOLD-OUT – EXPERIMENT / 2



- Clear pessimistic bias for small training sets – we learn a much worse model than with 500 observations.
- But increase in variance when test sets become smaller.

## BIAS-VARIANCE OF HOLD-OUT – EXPERIMENT / 3

- Let's now plot the MSE of the holdout estimator.
- NB: Not MSE of model, but squared difference between estimated holdout values and true performance (horiz. line in prev. plot).
- Best estimator is ca. train set ratio of 2/3.
- NB: This is a single experiment and not a scientific study, but this rule-of-thumb has also been validated in larger studies.

